DOCUMENT RESUME

ED 431 784 TM 029 864

AUTHOR Hohn, Alisabeth M.; Veitch, William R.
TITLE The Colorado Basic Literacy Act: Multiple Measures in Action.
PUB DATE 1999-04-00
NOTE 10p.; Paper presented at the Annual Meeting of the American Educational Research Association (Montreal, Quebec, Canada, April 19-23, 1999).
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Analysis; *Elementary School Students; Evaluation Methods; Grade 3; *Low Achievement; *Measurement Techniques; Performance Based Assessment; Primary Education; Reading Achievement; State Legislation
IDENTIFIERS *Basic Literacy Act (Colorado 1996); *Multiple Measures Approach

ABSTRACT

The Colorado Basic Literacy Law, enacted in 1996, requires that school districts report the number and percentage of students who are reading at or above grade level in grade 3, are on individualized literacy plans, or improve their reading achievement by two or more grade levels in a single year. The law requires the use of multiple indicators or a "body of evidence" for the first two of the requirements. This paper describes the impact of the new law on Colorado school districts and the multiple measures districts are using. The focus is on the first requirement, the number and percentages of third graders reading at or above the third grade level. The decision is based on standardized test results from an individual reading inventory and the state's third-grade reading test. A third indicator may be added, chosen from the state's approved list. Each district must set performance expectations for each instrument, and each must determine how to combine the evidence from multiple measures. As procedures now stand, information from the various districts will not be comparable because districts set their own cut points and make their own combining rules. An attachment lists the third grade reading proficiencies. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 431 784

The Colorado Basic Literacy Act: Multiple Measures in Action

AERA Division H Symposium on Multiple Measures

By

Alisabeth M. Hohn, Ph.D. & William R. Veitch, Ph.D.

Department of Planning, Research, and Evaluation
Colorado Springs School District Eleven

A. Hohn

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

- ☐ This document has been reproduced as received from the person or organization originating it.
- ☐ Minor changes have been made to improve reproduction quality.
- • Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

April 1999
Montreal, Canada

TM029864

Address information requests to:

Alisabeth M. Hohn, Ph.D.
Assessment Coordinator
Colorado Springs School District Eleven
1115 N. El Paso St.
Colorado Springs, CO 80903

ahohn@cssd11.k12.co.us

BEST COPY AVAILABLE

Introduction

In 1996, the Colorado Basic Literacy Act (CBLA) was signed into law. Beginning in the spring of 1999, school districts are required to report the number and percentage of students who:

1. are reading at or above grade level (grade 3)
2. are on Individualized Literacy Plans (ILPs)[1] (grades K-3 and adding one grade level per year thereafter)
3. improve their reading achievement by two or more grade levels within a single year of instruction (grades K-3)

The law requires the use of multiple indicators or a "body of evidence" for the first two of the three requirements above.

In this symposium, the multiple measures issue is being discussed from several perspectives. In order to understand *our* approach to this problem and its associated challenges and decisions, some extra background information will be useful.

As with most new laws, there appear to be benefits from and drawbacks to the CBLA. Some benefits are:

- Updated training for all K-3 teachers in reading
- Accountability in K-2
- Colorado Department of Education (CDE) provided a list of literacy proficiency expectations in grades K, 1, 2, and 3
- Improved reading instruction and progress monitoring for *all* children
- Improved communication with parents

Some general challenges are:

- Resources (unfunded mandate)
- Training
- Time for teachers to administer assessments and interpret results

More specifically, from the perspective of a school district assessment unit, the assessment and measurement challenges posed by this new law are numerous and substantial. Some of these issues are:

1. Conflict of purposes (instruction and large-scale accountability)
2. Data management and logistics
3. Standardization (district-wide instruments, classroom evidence)
4. Multiple measures – which ones to employ? For which purposes?
5. Setting performance expectations on each instrument
6. Combining rules for evidence from multiple measures
7. What constitutes growth?

---

[1] An ILP is written for any student who is reading below grade level. Its main purposes are to provide the student with additional reading instruction and to formally involve parents in their child's literacy instruction. The ILP must be revisited every six months until the child has "caught up" to grade level.

8. Desire for predictive validity (before data are available)
9. Comparability of data from different districts
10. Lack of measurement expertise in many Colorado school districts

To keep the issues manageable in the space available, this presentation focuses on the first of the three reporting requirements: the number and percentage of 3rd graders reading at or above the third grade level. Several of the challenging issues listed above come into play. This presentation focuses on the following:

- Choice of measures
- Setting performance expectations on each instrument
- Combining rules for evidence from multiple measures

One more challenge that impacts these decisions is *consequences*. A child who is not reading on grade level by the end of third grade is placed on an ILP and cannot move on to 4th grade reading instruction until he or she is ready. As the law is presently written, the student is not retained, but we anticipate a social stigma for these students. Because of this *individual level of accountability*, it is important to correctly classify as many students as possible. Aggregate results are also important. School and district results become public information that probably will be reported in our local newspapers. These results also contribute to each school's yearly accreditation report. Our decisions must be credible for individual students, schools, and the public.

Choice of Measures

Until the spring of third grade, the decision about whether a student is reading on grade level is based on the results from an individualized reading inventory and classroom evidence. Teacher judgment based on documented evidence is the driving force behind the decision. In contrast, the decision about a third grader's reading level is based solely on standardized test results: an individual reading inventory and results from the state-wide, third grade reading test. Districts may choose a third indicator, but it must come from the state's approved list.

Individual Reading Inventory
The individual reading inventory is a required element. Districts may choose from:

- Qualitative Reading Inventory (QRI-II)
- Flynt Cooter
- Basic Reading Inventory (Johns)
- Running Record with Comprehension or Retelling (Celebration Press, Wright Group)
- District developed assessment with researched and documented results

The curriculum department from our district chose to administer the QRI-II (Leslie & Caldwell, 1995). The resulting score is a student's "instructional reading level," not to be confused with grade levels. The QRI-II consists of a series of increasingly difficult reading passages in narrative and expository genre ranging from preprimer to junior high level. Not surprisingly,

recent research has shown that students read approximately one level lower on the expository passages than on the narrative passages (Felknor, 1999). For the sake of time, in District Eleven, the curriculum department chose to administer the series of narrative passages only.

State Assessment

Results from the grade three Colorado Student Assessment Program Test (CSAP) also are a required element. The test was administered for the first time last year, with no stakes attached. This year, the results count. In March, students complete a two-period, mixed-format test of reading comprehension. Results are returned in May and are reported in "proficiency levels." Each student's performance is categorized as Advanced, Proficient, Partially Proficient, or Unsatisfactory. For a student to have "passed" the CSAP test, he or she must receive a proficient or advanced rating.

Third Indicator

School districts may include a third indicator if they choose. Approved instruments are:

- Reading series assessments (e.g., Houghton-Mifflin Invitations to Literacy)
- ITBS with Constructed Response or Integrated Performance Assessments
- Northwest Evaluation Association (NWEA) Levels Tests
- Terra Nova (CTB)

District Eleven has been using NWEA Levels Tests in grades 3-8 for many years. Levels Tests are administered to all students in reading, language, and math every fall and spring. We chose to include the spring 3rd grade reading score as a third indicator.

All three of the measures we selected map nicely onto the descriptors of what children are supposed to know and be able to do by the end of 3rd grade (attached).

Setting Performance Expectations on each Instrument

CSAP

CSAP is already reported in proficiency levels. In order to set proficiency levels on the other two tests, we needed the CSAP scores and the narrative descriptors of what performance "looks like" at each proficiency level. By winter 1999, we had the information needed.

The CSAP test score is the anchor against which performance expectations on the other instruments were set. While the CSAP score is technically part of the "body of evidence" used to decide if a student is reading at or above the third grade level, both the CSAP score distribution and the reports submitted to the state are public documents. If the state test results say that 50% of third graders are proficient or above, and the reports we send back to the state say that 85% of our students are proficient or above, the public (especially the newspapers) are likely to think that we are systematically inflating our scores. In other words, the distribution we report to the state needs to look a lot like the CSAP score distribution. Given this set of requirements, it is important that the performance levels for the other two indicators are equally rigorous.

QRI-II

This fall (1998) was the first time that we administered the QRI-II district wide, so there were no within-district spring data with which to set performance expectations. Luckily, a study by Catherine Felknor (1999) in the Denver Metro area districts compared CSAP and QRI-II scores. She found that performance on the QRI-II narrative passages was more highly correlated with CSAP scores than was performance on the QRI-II expository passages. This was good news, as we were considering switching to the expository passages, and were not interested in doubling administration time by giving both types of passages to each student, as is the policy in some districts.

Her results showed that a proficient student on the CSAP most likely had an instructional reading level at passage 4. The advanced student was reading passages 5 or above at an instructional level. A partially proficient reader could read passage 3, and an unsatisfactory student was reading passage 2 or below at an instructional level.

Levels Tests

Our NWEA Levels Test scores are reported in Rasch Units, or RITs. A RIT is a scale score based on a one-parameter item response model ranging from approximately 150-270 encompassing achievement from grades 3-8. Students take narrow band tests according to their ability level (as opposed to grade level), so the test does a better job of measuring student performance in the tails of the distribution than do typical standardized tests.

Until now, we have reported student scores only in RITs and local percentiles, not in terms of proficiency levels. Adding descriptive labels to the scale should help teachers and parents more effectively interpret their children's performance.

In the spring of 1997, we used a student-centered cut score procedure called the borderline method (Livingston & Zieky, 1982) to set some preliminary proficiency levels on the test. We never published the results because teachers set their expectations quite low. At that point in time, we had not administered a CSAP test in third grade, and did not expect our state test results to be so favorable. A paper describing our findings was presented at last year's AERA conference (Veitch, 1998).

In February 1999, we tried again, this time with one year of CSAP results and the associated proficiency descriptors in hand. Setting cut scores on our Levels Tests involved two steps. First, we had last year's Levels Test scores and CSAP scores for the same students. We used an equipercentile distribution to translate CSAP scores into RIT scores. So, if 60% of our district students were proficient or above on the CSAP, we could use our local Levels Test norms to determine the associated RIT score.

Next, we arranged for Dan Lewis from CTB/McGraw-Hill to facilitate a bookmarking session using our levels test items and the state's proficiency descriptors. Colorado uses the bookmarking procedure to set proficiency levels on our state tests. Without seeing our equipercentile distribution, the teachers who participated in the bookmarking workshop independently recommended the same cutpoints that we did. Further, our teachers found the procedure worthwhile and it validated our use of Levels Tests for this purpose.

Combining Rules for Evidence from Multiple Measures

If the distribution of students in the four proficiency categories ultimately must look like the distribution of scores from the state test, then why bother to use a body of evidence in the first place? In our search for literature on multiple measures, we found many authors who explain why using multiple measures is typically a good thing to do (decrease in error, increased score stability, allows a student to have a bad day, etc.) (Crone, Lang, Teddlie, & Franklin, 1995, Linn, 1998), but few actually do it, and even fewer write thoughtfully about their process and justifications for their decisions. We are pleased to see that CRESST is working to write guidelines people might follow when making these kinds of decisions in the future.

We have been wrestling with the concept and realities of multiple measures since the CBLA was enacted in 1996. In a 1997 presentation to the Association of Colorado Education Evaluators, Bob Linn suggested a process of careful instrument selection and cut scoring, followed by a judgment-based weighting and combining process. Gene Adcock from Prince George's County introduced us to value-added models using HLM, but we found those difficult to explain and inappropriate for the multiple levels of accountability required in this situation. At the CRESST conference in September, several California school districts discussed their multiple indicator systems for accreditation (e.g., Long Beach Unified School District, 1998).

After considering these and other ideas, the "KISS Rule" prevailed. Because our community is suspicious of anything the school district does, and our teachers will be the messengers, we decided that it was in our best interest to *keep it simple*. We are using three indicators, and have confidence in the methods used to derive proficiency levels on the instruments. We decided that if a student can demonstrate proficiency on two of the three standardized instruments, then we are comfortable saying that he/she is reading at or above the third grade level. Our teachers understand this, and can communicate it to students, parents, and to the public.

If a student is missing a CSAP score, the other two scores both must be proficient or above in order to say that the student is reading at or above grade level. If either of the other two scores is missing, the test can be made up at the school site.

Performance Expectations at the End of Third Grade

Proficiency Levels	QRI-II	DALT (RIT)	CSAP
Advanced	5 and above	220 and above	Adv. (561 and above)
Proficient	4	200-219	Prof. (482-560)
Partially Proficient	2 and 3	189-199	PP (445-481)
Unsatisfactory	1 and below	188 and below	Unsat. (444 and below)

BEST COPY AVAILABLE

After we receive a complete set of data this summer, we will revisit our cut score and combining rules decisions. We will perform a discriminant analysis and work to improve classification accuracy. This was the best we could do, given limited time and resources with incomplete data.

Conclusions

With the advent of the CBLA it is clear that early literacy is important in Colorado. Even before the reporting requirement has begun, we are seeing attention to early literacy teaching and learning in ways we have never seen before. In School District Eleven we are lucky to have the technical expertise to wrestle with the complex issues the law is creating. Only a handful of districts employ someone with measurement expertise, however. In this situation, ignorance must be bliss, because it keeps the rest of us up nights.

The bigger issues are ones for the legislature and State Department of Education to resolve. For reasons with which we are all familiar, using multiple measures can be a very good idea. However, with the CBLA, districts choose their own instruments, administer them under conditions standardized within their districts, set their own cut points, and make their own combining rules decisions. Equating studies of the instruments on the approved list have not been undertaken. Teachers are not equally well trained to administer the individual reading inventories. Districts without technical and literacy expertise will make arbitrary cut point and instrument combining decisions. In sum, despite the legislature's interest in ranking and comparing districts with one another using multiple measures, these data simply will not be comparable.

BEST COPY AVAILABLE

References

Crone, L. J., Lang, M. H., Teddlie, C., & Franklin, B. J. (1995). *Achievement measures of school effectiveness: Comparisons of model stability across years.* Applied Measurement in Education, 8, 353-365.

Felknor, C. (1999). [QRI-II Norming Study – Year 2]. Unpublished raw data.

Leslie, L., & Caldwell, J. (1995). *Qualitative Reading Inventory – II.* New York: Harper Collins.

Linn, R. (1998, September). *Accountability and Information Systems.* Paper presented at the CRESST Conference, Los Angeles, CA.

Livingston, S. A., & Zieky, M. J. (1982). Passing scores: A manual for setting standards of performance on educational and occupational tests. Princeton, NJ: Educational Testing Service.

Long Beach Unified School District (1998, September). *Long Beach Unified School District Standards-Based Assessment System,* Paper presented at the CRESST Conference, Los Angeles, CA.

Veitch, W. R. (1998, April). *Establishing Cut-Points on Local Achievement Assessments.* Paper presented at the American Educational Research Association, San Diego, CA.

BEST COPY AVAILABLE

Colorado Basic Literacy Act

3rd Grade Proficiencies

Understanding of Text

- Adjust reading pace to accommodate purpose, style, and difficulty of material.
- Summarize text passages
- Apply information and make connections from reading

Integration of Cueing Systems

- Use word attack skills to read new and unfamiliar words (graphophonics)
- Use sentence structure, paragraph structure, text organization, and word order (syntax)
- Use and apply background knowledge, experience, and context to construct a variety of meanings over developmentally appropriate texts (semantics)
- Use strategies of sampling, predicting, confirming, and self-correcting quickly, confidently, and independently (graphophonics, syntax, and semantics).